

Komparasi Tingkat Akurasi Random Forest dan C4.5 Pada Klasifikasi Performa Mahasiswa

Sudarto¹, Erlanie Sufarnap², Mirza Ilhami³

¹Universitas Mikroskil, sudarto@mikroskil.ac.id

²Universitas Mikroskil, erlanie@mikroskil.ac.id

³Universitas Mikroskil, mirza.ilhami@mikroskil.ac.id

Abstract

Mahasiswa merupakan salah satu faktor keberhasilan bagi Perguruan Tinggi Negeri (PTN) dan Perguruan Tinggi Swasta (PTS). Sehingga pengukuran performa mahasiswa tentunya sangat penting dilakukan sebagai bagian dalam melakukan pengawasan dan memahami peningkatan performa mahasiswa. Peningkatan performa mahasiswa dipengaruhi oleh beberapa faktor seperti ekonomi keluarga, pendidikan orang tua, media sosial, lingkungan, prestasi akademik. Penelitian ini bertujuan untuk membandingkan metode random forest dan metode C4.5 dalam melakukan prediksi dan klasifikasi dataset performa mahasiswa. Penelitian ini dilakukan mulai dari tahap pengumpulan data menggunakan data sekunder dari sumber repository UCI machine learning yang terdiri dari 145 data mahasiswa dan 33 atribut. Berikutnya dilakukan tahap preprocessing dengan melakukan seleksi fitur menggunakan information gain sehingga diperoleh 9 atribut yang berpengaruh dalam pengujian. Tahap pengujian dilakukan dengan split data terhadap data training dan data testing serta dengan menggunakan k-fold cross validation. Hasil dari pengujian menunjukkan bahwa nilai akurasi tertinggi yang diperoleh metode random forest sebesar 92,86 % melalui pengujian dengan split data 90:10. Untuk metode C4.5 diperoleh nilai akurasi tertinggi sebesar 75,34 % melalui pengujian dengan split data 90:10. Dari hasil pengujian kedua metode tersebut diperoleh nilai akurasi metode random forest lebih unggul 17,52 % dibandingkan metode C4.5 sehingga layak direkomendasikan dalam melakukan prediksi dan klasifikasi pada dataset performa mahasiswa.

Keywords: random forest, C4.5, Machine learning, performa mahasiswa, information gain

Corresponding Author: sudarto

Publication Date: April 1, 2025

Email: sudarto@mikroskil.ac.id

This is a Creative Commons Attribution-NonCommercial 4.0 International License (CC BY-NC 4.0)

Pendahuluan

Performa mahasiswa merupakan hasil dari kegiatan belajar mengajar yang telah diikuti oleh mahasiswa setelah masa perkuliahan dilaksanakan. Biasanya, pengukuran performa mahasiswa dilakukan dengan mengadakan ujian atau tugas berkelanjutan [1]. Hal terpenting yang perlu diperhatikan pada saat proses pembelajaran berlangsung adalah pengawasan mahasiswa yang masuk, peningkatan performa mahasiswa, prestasi yang dicapai mahasiswa, rasio jumlah mahasiswa yang lulus terhadap total mahasiswa dan kompetensi lulusan. Hasil capaian tersebut tentu saja berpengaruh pada ketepatan mahasiswa dalam menyelesaikan waktu studinya [2].

Faktor-faktor yang memengaruhi performa akademik mahasiswa sangat kompleks dan dapat dikelompokkan ke dalam tiga kategori utama:

Bagi Perguruan Tinggi Negeri (PTN) dan Perguruan Tinggi Swasta (PTS), mahasiswa merupakan salah satu faktor keberhasilan [3]. Faktor yang mempengaruhi performa akademik mahasiswa sangatlah kompleks mencakup yaitu personal dan media sosial, akademik, lingkungan, ekonomi dan penilaian internal. Rata-rata mahasiswa mengalami putus kuliah karena mengundurkan diri dan dikeluarkan. Faktor lainnya terkait personal dan sosial terhadap prestasi akademik pada 100 mahasiswa di sebuah Universitas di Hong Kong meliputi perilaku penggunaan media sosial, perhatian dan dukungan dari orang tua, kehadiran di kelas, beban tanggungan keluarga, pekerjaan orang tua, yang menunjukkan hasil hubungan positif antara tingginya penggunaan media sosial dengan tingkat prestasi akademik mahasiswa yang rendah [4]. Selain itu, faktor ekonomi orang tua mahasiswa serta tingkat pendidikan orang tua maupun faktor pribadi sangat mempengaruhi kinerja mahasiswa di institusi [5]. Faktor Akademik mencakup indeks prestasi kumulatif, nilai ujian masuk, hasil nilai mata kuliah, partisipasi dalam tugas proyek, dan kesiapan menghadapi ujian tengah atau akhir semester, yang merupakan bagian dari proses evaluasi berkelanjutan dari performa mahasiswa, ditemukan memiliki dampak tertinggi pada hasil semester akhir [6]. Mahasiswa yang lebih sering menggunakan sosial media dalam kesehariannya cenderung mengalami gangguan fokus belajar. Beberapa faktor kunci yang mempengaruhi kinerja akademik siswa. Faktor lainnya yang dianggap berpengaruh dalam performa mahasiswa yaitu mencakup indeks prestasi, pekerjaan dan tingkat pendidikan ayah, tingkat pendidikan ibu, kehadiran di kelas, jumlah tanggungan keluarga, biaya kuliah, nilai ujian sekolah menengah, pekerjaan ibu, jam belajar rata-rata, prestasi akademik, jalur masuk perguruan tinggi, rata-rata penggunaan ponsel cerdas untuk hiburan, jenis kelamin, pendapatan

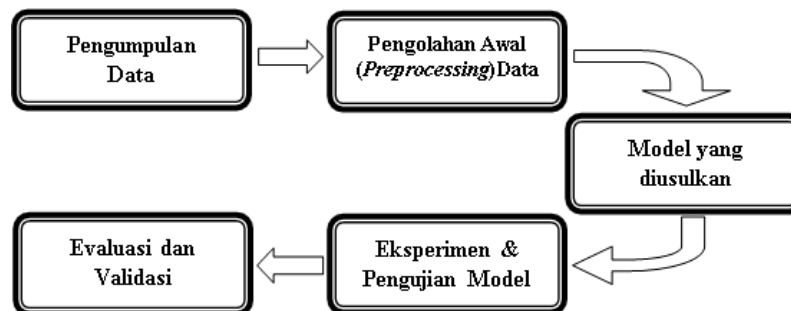
keluarga, jenis tempat tinggal, dan jarak tempat tinggal ke perguruan tinggi. Beberapa faktor seperti faktor organisasi dan nilai ujian akhir sekolah menengah pertama ditemukan memiliki pengaruh yang tidak signifikan terhadap kinerja akademik [7].

Penggunaan algoritma Random Forest yang dikombinasikan dengan pemilihan fitur Information Gain untuk memprediksi kinerja akademik mahasiswa di Universitas Sebelas Maret. Akurasi tinggi yang dicapai menggarisbawahi potensi model ini untuk identifikasi awal siswa yang berisiko kinerja akademik rendah, memungkinkan intervensi tepat waktu. Faktor-faktor berpengaruh yang diidentifikasi memberikan wawasan berharga bagi lembaga pendidikan untuk mendukung siswa mereka dengan lebih baik [7]. Dalam penelitian yang memprediksi kinerja akademik siswa, diperoleh hasil bahwa variabel yang berpengaruh penting yaitu jarak dan waktu tempuh ke sekolah. Didapatkan bahwa algoritma Random Forest lebih unggul dari SVM, kNN, Naïve Bayesian Classifier (NBC) [8]. Begitu juga penelitian mengenai prediksi kinerja siswa, dengan 300 data dan 24 variabel, hasil penelitiannya menunjukkan metode Random Forest memiliki tingkat akurasi yang lebih tinggi yaitu 99% dibandingkan dengan metode C4.5 yang hanya 74% [6]. Random Forest dapat memproses data yang tidak seimbang dan menangani pemasangan lebih efektif. Merujuk dari beberapa penelitian yang telah dilakukan menunjukkan bahwa prestasi akademik sangat dipengaruhi oleh beberapa faktor, antara lain faktor sosial, ekonomi dan lingkungan, maka faktor-faktor tersebut perlu dianalisis kembali kontribusi untuk memprediksi kinerja akademik mahasiswa. Dalam penelitian [9] bahwa tingkat akurasi metode C4.5 lebih unggul dibandingkan random forest sebesar 3,34 %. hasil dan analisis pengujian yang telah dilakukan [10] bahwa klasifikasi prediksi prestasi akademik mahasiswa menggunakan metode C4.5 menghasilkan akurasi 87.1 %, nilai presisi sebesar 85.4% dan nilai recall sebesar 87.1% , metode Random Forest menghasilkan akurasi sebesar 92.4%, nilai presisi sebesar 91.4% dan nilai recall sebesar 92.4%. Sehingga metode Random Forest memiliki akurasi, presisi dan recall yang lebih baik dibandingkan dengan metode C4.5.

Meskipun banyak penelitian telah menggunakan *machine learning* untuk prediksi performa akademik, belum banyak yang mengkaji perbandingan antara algoritma Random Forest dan C4.5 secara menyeluruh, khususnya dengan analisis *gain ratio* dan validasi silang (*cross-validation*) dalam konteks kampus Indonesia. Penelitian ini juga memperhatikan aspek *preprocessing* secara menyeluruh agar hasil prediksi memiliki reliabilitas yang lebih tinggi. Penelitian ini bertujuan untuk mengisi celah tersebut dengan melakukan analisis komparatif antara algoritma Random Forest dan C4.5 berdasarkan data dari UCI Machine Learning Repository.

Metode Penelitian

Metode yang digunakan dalam penelitian ini adalah eksperimental dan studi komparatif berbasis data mining yang menggunakan pendekatan kuantitatif untuk membandingkan dua algoritma *supervised learning*, yaitu Random Forest dan C4.5. Penelitian eksperimental bisa dilakukan pada dua kelompok dimana kelompok satu disebut kontrol yang tanpa diberi perlakuan apapun sedangkan kelompok kedua diberikan perlakuan [11]. Adapun tahapan dalam penelitian ini dapat dilihat pada gambar 1 dibawah ini :



Gambar 1. Tahapan Penelitian

Tahap pengumpulan dataset yang digunakan adalah data sekunder yaitu data kelulusan mahasiswa. Data sekunder merupakan data yang diperoleh secara tidak langsung bersumber dari dokumentasi, literatur, buku, jurnal dan informasi lainnya yang ada hubungannya dengan masalah yang diteliti. Data sekunder yang diambil pada penelitian ini adalah melalui website UCI Machine Learning <https://archive.ics.uci.edu/datasets>. Dataset yang digunakan dari Uci Machine Learning bertujuan untuk memprediksi performa mahasiswa yang terdiri dari 145 data mahasiswa dan 33 atribut. Untuk pengolahan data menggunakan menggunakan *software rapid miner versi 10.2* dan bahasa pemrograman *pyhton*.

Tahapan selanjutnya yaitu pengolahan awal (*preprocessing*) terhadap data. Pada tahap ini dilakukan proses pembersihan data dengan melakukan analisis dan penanganan *missing value*, *noise* maupun *outlier*. Hasil penanganan

dari *missing value*, *noise* maupun *outlier* dapat dilihat pada gambar 2 dibawah ini

```

TABEL 1: Penanganan Missing Values
Nama Kolom  Jumlah Missing  Metode Penanganan

Empty DataFrame
Columns: [Nama Kolom, Jumlah Missing, Metode Penanganan]
Index: []

TABEL 2: Penanganan Outlier (IQR)
Empty DataFrame
Columns: []
Index: []

TABEL 3: Penanganan Noise
Jenis Noise          Deskripsi Masalah          Tindakan
Nama kolom tidak rapi  Terdapat newline, spasi ganda, simbol  Normalisasi nama kolom
Data duplikat          0 baris duplikat ditemukan          Hapus duplikat

```

Gambar 2. Hasil Penanganan *Missing Value*, *Noise* dan *Outlier* pada *software python*

Berdasarkan gambar 2 bahwa hasil eksplorasi menunjukkan bahwa tidak terdapat *missing value* pada seluruh atribut dalam dataset. Hal ini mengindikasikan bahwa data yang digunakan telah memiliki kelengkapan (*completeness*) yang baik. Sehingga tidak diperlukan metode imputasi seperti mean, median, atau mode dan mengurangi potensi bias yang sering muncul akibat penanganan nilai hilang secara paksa. Begitu juga dalam penanganan *outlier* berdasarkan metode *Interquartile Range* (IQR), tidak ditemukan outlier pada atribut numerik. IQR digunakan untuk mengidentifikasi nilai ekstrem di luar rentang $Q1 - 1.5 \times IQR$ hingga $Q3 + 1.5 \times IQR$. Distribusi data cenderung simetris atau normal dan model *machine learning* yang sensitif terhadap *outlier* (seperti *Decision Tree*) dapat bekerja lebih stabil. Dalam penanganan noise telah dilakukan normalisasi terhadap beberapa atribut dan tidak ditemukan adanya duplikasi data. Dari hasil penanganan missing value, outlier dan noise yang dilakukan ini menunjukkan bahwa dataset yang digunakan tidak memiliki missing value dan outlier serta struktur kolom yang bersih, terstandarisasi maupun bebas duplikasi sehingga dataset memiliki kualitas data yang tinggi.

Selanjutnya dilakukan seleksi fitur dari data yang telah didapat menggunakan metode *information gain*. Atribut data yang telah ditentukan berdasarkan faktor-faktor pada dataset tadi akan direduksi dan diseleksi untuk mengurangi atribut yang kurang relevan atau kurang berpengaruh terhadap proses pembelajaran data. Dari total 145 data dan 33 atribut akan dilakukan seleksi fitur menggunakan *information gain* dengan meranking nilai gain dari setiap atribut yang paling berpengaruh sehingga diperoleh hasil akhir perankingan atribut seperti pada Tabel 1

Tabel 1. Hasil seleksi atribut

No	Atribut	Keterangan	Gain Ratio
1.	Scholarship type:	1: None, 2: 25%, 3: 50%, 4: 75%, 5: Full	2,501121098
2.	Grade	0:Fail, 1:DD, 2:DC, 3:CC, 4:CB, 5:BB, 6:BA, 7:AA	0,720932407
3.	Cumulative grade point average in last semester	1: <2.00, 2: 2.00-2.49, 3: 2.50-2.99, 4: 3.00-3.49, 5: above 3.49	0,64221653
4.	Course ID		0,272800494
5.	Impact of your project/activities on your success	1: married, 2: divorced, 3: died - one of them or both	0,309671932
6.	Expected cumulative grade point average in last semester	11: <2.00, 2: 2.00-2.49, 3: 2.50-2.99, 4: 3.00-3.49, 5: above 3.49	0,305896915
7.	Sex	1: female, 2: male	0,165104367
8.	Preparation to midterm exams 2	1: closest date to the exam, 2: regularly during the semester, 3: never	0,114020469
9.	Attendance to the seminars/conferences)	1: Yes, 2: No	0,113626777

Model yang diusulkan pada penelitian ini adalah menggunakan algoritma klasifikasi data mining decision tree Random Forest dan C4.5. setelah data diolah dan menghasilkan model, maka dilanjutkan dengan pengujian menggunakan *k-fold cross validation*, kemudian dilakukan evaluasi hasil dengan *confusion matrix* untuk mengukur akurasi, presisi dan recal. Penilaian terhadap hasil statistik analisis akan menggeneralisasi kumpulan data independen dapat dilakukan dengan penggunaan model teknik validasi yang disebut dengan *cross validation*. Tujuan dari penggunaan teknik tersebut, yakni untuk prediksi model dan prediksi tingkat keakuratan sebuah model prediktif ketika digunakan. Bias yang terdapat pada data diketahui dapat dihilangkan melalui penggunaan *K-fold cross validation*. Pada tahap ini, model yang terbentuk akan diuji dengan *confusion matrix* untuk mengetahui tingkat akurasi. Diketahui bahwa, pada sistem terdapat sebuah kemampuan dan kemampuan yang berfungsi untuk menemukan peringkat yang paling relevan dapat dievaluasi melalui penggunaan hasil perhitungan dari *precision*. Selain itu, kemampuan pada sistem yang berfungsi untuk menemukan semua item yang relevan dari koleksi dokumen dapat dievaluasi melalui penggunaan hasil perhitungan dari *recall*. Validasi hasil penelitian dilakukan dengan mengambil sampel sebanyak 145, kemudian data diuji dengan metode decision tree Random Forest dan C4.5 menggunakan *10-fold cross validation*.

Hasil Penelitian

Pada tahap pengujian dan eksperimen klasifikasi data, akan dilakukan pembangunan model menggunakan algoritma *decision tree* random forest dan C4.5. Berikut hasil akurasi sebelum dilakukan seleksi fitur dari metode *decision tree* C4.5 pada tabel 2 dibawah ini :

Tabel 2. Nilai akurasi C4.5 sebelum dilakukan seleksi fitur.

Algoritma	Akurasi
C4.5 (sebelum di seleksi)	73,33%

Pada metode random forest, berdasarkan percobaan jumlah pohon yang dilakukan seperti yang tersaji pada tabel 3 dengan variasi jumlah pohon 10, 50, 100, 500, dan 1000.

Tabel 3. Nilai akurasi random forest sebelum seleksi fitur

Jumlah Pohon	Akurasi		
	90:10	80:20	70:30
10	85.71%	82.76%	84.09%
50	85.71%	82.76%	86.36%
100	78.57%	82.76%	86.36%
500	85.71%	82.76%	86.36%
1000	85.71%	82.76%	86.36%

Pada tabel 3 didapatkan hasil bahwa akurasi tertinggi mulai didapat pada split data 30 : 70 dengan tingkat akurasi 86.36%. Peneliti memilih 1000 pohon karena percobaan yang dilakukan pada sistem juga semakin banyak, sehingga pada percobaan seterusnya akan menggunakan 1000 jumlah pohon dalam proses klasifikasi.

Selanjutnya dilakukan pengujian dengan penggunaan *K-fold cross validation* pada metode random forest dan C4.5 sebelum dilakukan seleksi fitur dan setelah seleksi fitur dengan *information gain*. Hasil pengujian dapat dilihat pada tabel 4 dan tabel 5 dibawah ini :

Tabel 4. Hasil Pengujian Metode C4.5

Eksperimen	Akurasi	Presisi	Recall
10 K-fold dan confidence 0,25 (sebelum seleksi fitur)	73.89%	87.35%	91.58%
	73.46%	86.80%	90.45%

20 K-fold dan confidence 0,25 (sebelum seleksi fitur)			
30 K-fold dan confidence 0,25 (sebelum seleksi fitur)	73.96%	87.87%	90.13%
10 K-fold dan confidence 0,5 (setelah seleksi fitur)	74.00%	87.61%	91.72%
20 K-fold dan confidence 0,5 (setelah seleksi fitur)	73.74%	87.69%	90.77%
30 K-fold dan confidence 0,5 (setelah seleksi fitur)	74.01%	87.87%	90.18%
90% : 10%	75.34%	82.26%	91.86%
80% : 20%	74.46%	80.08%	90.05%
70% : 30%	73.47%	77.97%	89.59%

Tabel 5. Hasil Pengujian Random Forest

Eksperimen		Akurasi	Presisi	Recall
Sebelum seleksi fitur	90:10	85.71%	75.00%	100.00%
	80:20	82.76%	70.59%	100.00%
	70:30	86.36%	75.00%	100.00%
Sesudah seleksi fitur	90:10	92.86%	100.00%	83.33%
	80:20	82.76%	70.59%	100.00%
	70:30	86.36%	75.00%	100.00%
Sebelum seleksi fitur	5-folds cross validation	91.72%	87.50%	93.33%
	10-folds cross validation	90.34%	87.10%	90.00%
Sesudah seleksi fitur	5-folds cross validation	91.72%	88.71%	91.67%
	10-folds cross validation	92.41%	90.16%	91.67%

Pada tabel 4 dan tabel 5 dapat dilihat bahwa tingkat akurasi tertinggi metode C4.5 sebesar 75,34 % dan random forest sebesar 92,86 % melalui eksperimen dengan penggunaan split data 90:10 setelah seleksi fitur. Ini menunjukkan bahwa tingkat akurasi metode random forest lebih baik dibandingkan metode C4.5 dalam melakukan klasifikasi performa mahasiswa. Hasil pengujian ini sejalan dengan penelitian yang dilakukan oleh [12] dan penelitian [13] bahwa metode random forest lebih dan lebih efektif dibandingkan metode C4.5 dalam memprediksi akurasi performa mahasiswa.

Pembahasan

Dalam penelitian ini dilakukan analisis menggunakan metode klasifikasi random forest dan C4.5 Setelah dilakukan seleksi fitur, selanjutnya terhadap model yang sudah dihasilkan tersebut dilakukan pengujian menggunakan split data antara data training dan data testing yaitu 90 : 10, 80 : 20, 70 : 30 dan mengulang pengujian tersebut beberapa kali. Dalam penelitian ini, proses eksperimen dan pengujian model menggunakan 1000 pohon. Nilai akurasi yang diperoleh dengan menggunakan algoritma random forest serta perbandingan data testing 10% : data training 90% didapatkan nilai akurasi = 92.86%, Nilai akurasi yang diperoleh dengan menggunakan metode random forest serta

perbandingan data testing 20% : data training 80% didapatkan nilai akurasi = 82.76% , Nilai akurasi yang diperoleh dengan menggunakan metode random forest serta perbandingan data testing 30% : data training 70% didapatkan nilai akurasi = 86.36%. Selanjutnya terhadap model yang sudah dihasilkan tersebut dilakukan pengujian menggunakan *k-fold cross validation* terdapat perbandingan nilai akurasi, presisi, *recall* dengan perbandingan 5-folds dan 10-folds cross validation. hasil dari *k-folds cross validation* didapatkan akurasi tertinggi didapatkan pada fold 10 sesudah seleksi fitur dengan nilai akurasi 92.41%. Adapun akurasi terendah terdapat pada fold 10 sebelum seleksi fitur dengan nilai akurasi 90.34%.

Hasil pengujian metode C4.5 dengan split data menggunakan data training dan data testing 90% : 10% mendapatkan nilai akurasi sebesar 75,34%, untuk data training dan data testing 80% : 20% mendapatkan nilai akurasi sebesar 74,46%, dan data training dan data testing 70% : 30% mendapatkan nilai akurasi sebesar 73,47%. Dari pengujian tersebut, terlihat dengan split data menggunakan data training dan data testing 90% : 10% menghasilkan akurasi terbesar yaitu 75,34%. Berikutnya, sebelum dilakukan pengujian menggunakan *k-fold cross validation* dengan data setelah seleksi fitur menghasilkan akurasi 73,22% dan akurasi mengalami peningkatan setelah menggunakan pengujian 10-fold cross validation dengan menggunakan setelah seleksi fitur menghasilkan akurasi 73,62%. Pengujian 10-fold cross validation dengan menggunakan data sebelum seleksi fitur menghasilkan akurasi 72,99% sedangkan dengan menggunakan data setelah seleksi fitur menghasilkan akurasi 73,62%. Hal ini menunjukkan adanya kecenderungan jika jumlah *X-fold cross-validation* semakin besar dan dikombinasikan dengan seleksi fitur maka kinerja nilai akurasi dan nilai presisi metode C4.5 juga meningkat [14].

Berdasarkan hasil pengujian, random forest secara konsisten menghasilkan akurasi yang lebih tinggi dibandingkan C4.5. Keunggulan ini berasal dari *ensemble learning* yang mengurangi *overfitting*, mampu menangani multikolinearitas, tahan terhadap noise dalam data [5] [12]. Dengan 1000 pohon, kestabilan model random forest meningkat [15]. Kinerja random forest yang lebih tinggi kemungkinan dipengaruhi oleh sifat *ensemble* yang mampu mengurangi *overfitting* dan mengatasi multikolinearitas, dibandingkan C4.5 yang lebih sensitif terhadap split data tertentu.

Kesimpulan

Berdasarkan hasil penelitian yang telah dilakukan diperoleh kesimpulan bahwa terdapat hasil akurasi yang berbeda pada metode random forest dan metode C4.5 dalam melakukan klasifikasi pada data performa mahasiswa. Perbandingan antara metode random forest dan C4.5 menunjukkan bahwa kedua metode memiliki selisih akurasi sebesar 17,52 %. Untuk metode random forest nilai akurasi tertinggi sebesar 92,86 % dan metode C4.5 dengan nilai akurasi tertinggi sebesar 75,34 % yang keduanya diperoleh melalui split data 90:10 setelah dilakukan seleksi fitur. Hasil dalam penelitian ini menunjukkan metode random forest lebih layak direkomendasikan dalam melakukan klasifikasi dataset performa mahasiswa dan mendeteksi risiko kegagalan studi mahasiswa lebih awal. Untuk penelitian selanjutnya dapat menggunakan data longitudinal, dan menggabungkan algoritma *hybrid* atau *deep learning* untuk peningkatan performa klasifikasi.

References

- [1] F. A. Saputra, I. Gartina Husein, and M. Qana'a, "Perancangan Dashboard untuk Monitoring Performa Mahasiswa D3 Sistem Informasi Fakultas Ilmu Terapan Berbasis Web," *Proceeding of Applied Science*, vol. 7, no. 6, p. 2713, 2021.
- [2] E. Prasetyowati, dan Nilam Ramadhani, J. Raya Panglegur Km, and J. Timur, "Sistem Evaluasi dan Klasifikasi Kinerja Akademik Mahasiswa Universitas Madura Menggunakan Naive Bayes dengan Dirichlet Smoothing," 2018.
- [3] E. Novianto, S. Suhirman, and D. Prasetyo, "Perbandingan Metode Klasifikasi Random Forest dan Support Vector Machine dalam Memprediksi Capaian Studi Mahasiswa," *JUPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)*, vol. 9, no. 4, pp. 1821–1833, Nov. 2024, doi: 10.29100/jupi.v9i4.5423.
- [4] S. Feng, Y. K. Wong, L. Y. Wong, and L. Hossain, "The Internet and Facebook Usage on Academic Distraction of College Students," *Comput Educ*, vol. 134, pp. 41–49, Jun. 2019, doi: 10.1016/j.compedu.2019.02.005.
- [5] S. Jayaprakash, S. Krishnan, and J. V, "Predicting Students Academic Performance using an Improved Random Forest Classifier," in *International Conference on Emerging Smart Computing and Informatics (ESCI)*, Institute of Information Technology, India, Mar. 2020, pp. 238–243.
- [6] S. Hussain, N. A. Dahan, F. M. Ba-Alwib, and N. Ribata, "Educational data mining and analysis of

- students' academic performance using WEKA," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 9, no. 2, pp. 447–459, Feb. 2018, doi: 10.11591/ijeecs.v9.i2.pp447-459.
- [7] M. Gusnina, Wiharto, and U. Salamah, "Student Performance Prediction in Sebelas Maret University Based on the Random Forest Algorithm," *Ingenierie des Systemes d'Information*, vol. 27, no. 3, pp. 495–501, Jun. 2022, doi: 10.18280/isi.270317.
- [8] D. Sharma and N. Agrawal, "Development of Modified CNN Algorithm for Agriculture Product: A Research Review," *Jurnal Ilmiah Teknik Elektro Komputer dan Informatika*, vol. 8, no. 1, p. 167, May 2022, doi: 10.26555/jiteki.v8i1.23722.
- [9] M. Azhari, Z. Situmorang, and R. Rosnelly, "Perbandingan Akurasi, Recall, dan Presisi Klasifikasi pada Algoritma C4.5, Random Forest, SVM dan Naive Bayes," *JURNAL MEDIA INFORMATIKA BUDIDARMA*, vol. 5, no. 2, p. 640, Apr. 2021, doi: 10.30865/mib.v5i2.2937.
- [10] S. Linawati, S. Nurdiani, K. Handayani, and Latifah, "PREDIKSI PRESTASI AKADEMIK MAHASISWA MENGGUNAKAN ALGORITMA RANDOM FOREST DAN C4.5," *Jurnal Khatulistiwa Informatika*, vol. VIII, no. 1, pp. 47–52, Jun. 2020, [Online]. Available: www.bsi.ac.id
- [11] Z. A. Hasibuan, "Metodologi Penelitian Pada Bidang Ilmu Komputer dan Teknologi Informasi," Fakultas Ilmu Komputer Universitas Indonesia, 2007.
- [12] A. Prabowo, S. Wardani, R. Wijaya Dewantoro, and W. Wesly, "Komparasi Tingkat Akurasi Random Forest dan Decision Tree C4.5 Pada Klasifikasi Data Penyakit Infertilitas," *Media Online*, vol. 4, no. 1, pp. 218–224, 2023, doi: 10.30865/klik.v4i1.1115.
- [13] J. Justam, N. Jamilah, S. M. Umar, E. Erlita, and J. Ramba, "Penerapan Algoritma C4.5 dan Random Forest untuk Pemetaan Kerusakan Jalan dengan WebGIS," *Jurnal Ilmiah Sistem Informasi dan Teknik Informatika (JISTI)*, vol. 7, no. 2, pp. 326–339, Oct. 2024, doi: 10.57093/jisti.v7i2.270.
- [14] Sudarto and E. Sufarnap, "Analisis Seleksi Fitur dengan Menggunakan Klasifikasi C4.5 dan Density Based Feature Selection (DBFS) dalam Memprediksi Kelulusan Mahasiswa," in *CITISEE*, Universitas AMIKOM Purwokerto, 2018.
- [15] P. Ayu Firnanda *et al.*, "Analisis Perbandingan Decision Tree dan Random Forest dalam Klasifikasi Penjualan Produk pada Supermarket," *Emerging Statistics and Data Science Journal*, vol. 3, no. 1, 2025.